7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Interactive auralization of self-generated oral sounds in virtual acoustic environments for research in human echolocation

David Pelegrin Garcia, Monika Rychtáriková, Christ Glorieux
Lab. Acoustics and Thermal Physics, Katholieke Universiteit Leuven, Heverlee, Belgium.

Brian F.G. Katz
Audio and Acoustics Group, LIMSI-CNRS, Orsay, France.

Summary
Some blind people are able to use self-generated oral sounds to detect the spatial configuration of the surroundings by listening to the sound patterns resulting from the combination of the direct sound and its reflections, applying the same principles as bats do in echolocation. We describe an auralization system that evokes passive virtual acoustic environments, i.e. where the sounds presented to users correspond to their self-generated oral sounds reflected/scattered by virtual boundaries – in real time – so that they have the impression of actually being surrounded by the simulated environment. This headphone-based system is responsive to head rotations in azimuth by using an orientation sensor. The propagation of sound between the mouth and the ears of a person inside a room, at one position and orientation, is characterized by means of the oral-binaural impulse response (OBRIR). Using commercial geometrical-acoustics software, we calculate offline the OBRIRs for all orientations at 15 degrees intervals. Low-latency convolution between the voice of the user picked with a microphone and the OBRIRs is performed with the software Max/MSP, which allows for real-time performance. Depending on the orientation given by the head tracker, Max/MSP interpolates the appropriate convolution outputs and delivers the resulting sound via headphones. This system will be used to examine the acquisition of spatial knowledge in new environments by means of echolocation, with a special focus on blind people.

PACS no. 43.55.-n, 43.38.-p

1. Introduction

Human beings typically rely on visual information to perform many tasks in everyday life, e.g. reading, writing or orienting oneself in an environment. According to the World Health Organization, blindness reduces people's ability to move unaided unless properly trained [1], and therefore it constitutes a barrier for the integration of blind people in society.

Currently, there are a number of blind (and sighted) people who, by clicking and listening to the echoes, are able to analyze and understand the spatial configuration of their surroundings. This technique is called echolocation and it is also used by different animal species (e.g. cetaceans and bats) [2]. Despite the benefits for blind people, only a minority of them use echolocation [3]. A lot of work has been done by the expert echolocator Daniel Kish and his collaborators to train and popularize this skill [4].

Echolocation in humans (a recent review of which is found in [5]) is most sensitive in the frontal direction, where hearing is most sensitive and voice directivity has its maximum, and provides a person with spatial information about distance, size and texture of an object, depending on the delay, intensity, and frequency filtering of the reflection – relative to the direct sound. Moreover, reflections from off-axis objects produce binaural cues [6]. This information gives echolocators (i.e. people who echolocate) valuable information for effective navigation and recognition of an environment [3]. Effective navigation, however, requires the combination of echolocation with other techniques like the white cane, more suited to detect obstacles that are close to the ground.

The major obstacles in learning echolocation are the amount of practice it requires and the cognitive load it involves, which makes it difficult to combine with other techniques. In addition, echolocation is

barely starting to be taught in a few countries; training methods are still in an early stage of development.

According to the current understanding, sensing via echolocation is built upon auditory mechanisms [5]. In consequence, the oral self-generated sound at the ears of an echolocator is characterized by the body-conducted sound, which propagates internally through the skull and body tissues into the cochlea, the airborne direct sound, travelling in the air between the mouth and the ears diffracted by the head, and the airborne reflected sound, which reaches the ears after being reflected or scattered at boundaries. The airborne direct and reflected sound propagation paths, independently of the source signal, are characterized with an impulse response sometimes referred to as Oral-Binaural Room Impulse Response (OBRIR) [7]. When echolocation is performed in different environments, only the airborne reflected sound propagation path changes.

In our present research, we have developed a head-tracked, headphone-based auralization system that is able to simulate passive virtual acoustic environments (VAEs); i.e. it produces the airborne reflected sound patterns that an actual object or room would produce in response to the oral sounds generated by a user. With such a system, we aim at

- Understanding the functioning of echolocation in conditions of variable reverberance and background noise.
- Assisting in the training of echolocation and evaluate different training strategies.

Such a system works by continuously convolving the sound produced by the user with the OBRIR corresponding to the current orientation and chosen environment. Because the user hears himself when producing sounds, the system omits the direct sound contribution and plays back only the reflected sounds. In order to achieve our goals, our system requires low latency in the audio chain, which allows simulating close boundaries.

Similar systems have been developed by Pörschmann [8] and Wefers [9] in the context of virtual reality, Picinali et al [10] in the context of virtual architectural navigation by the blind, Yadav et al [11] in the context of musical acoustics and Pelegrin-Garcia [12] for the study of preferred acoustic conditions for speaking in classrooms.

2. System description

2.1. Overview

The functional block diagram of the system for interactive auralization of self-generated oral sounds is shown in Fig. 1.

Given the description of a scene, defined by the geometry of the environment, the placement and the characteristics of source and receiver, acoustic simulations are performed with CATT Acoustic v9.0c. The result is an OBRIR that contains the acoustic information of the propagation of sound emanating from the mouth, interacting with the room, and returning to the ears. This OBRIR is post-processed to remove the simulated direct sound path, which in the architectural acoustics simulation software is not able to correctly model the near-field propagation of the sound from the mouth to the ears. This real acoustic path is also present in the VAE if open headphones are employed. Additional early delay is also removed from the response to compensate for the latency introduced later by the real-time processing. A single OBRIR corresponds to a single position in a single environment at a discrete orientation. An interactive VAE requires a set of OBRIRs derived at regular angles. We used 15° spacing to cover the possible orientations of the user in the 360° of the horizontal plane.

In an anechoic chamber a user wears Sennheiser HD570 open headphones with an orientation tracker XSens MTi mounted on top. His oral sounds are picked up with an omnidirectional Electret condenser microphone Sennheiser MKE-2P positioned 3 cm in front of the mouth. The microphone is equipped with a foam windscreen to minimize popping sounds due to turbulent airflow near the mouth. The signal from the microphone is split with one copy stream sent to the 2x31 one-third octave band equalizer DAP Audio DEQ-231 which is used to restore the natural quality of the airborne direct sound propagation between the mouth and the ears that is modified due to the presence of headphones on the ears. The second copy stream from the microphone is sent to an audio interface RME Fireface UCX, connected via USB to a computer running Max 6.1.6 under OS X v10.9. All audio processing is done at a sampling rate of 96000 Hz, with an I/O and signal vector size of 64 samples.

In Max, the user can select one of the pre-computed VAEs. When one is chosen, the OBRIRs corresponding to the different head orientations are loaded into a buffer, which is then linked to the convolution engine. The convolution engine was designed to run 48 parallel convolutions (24 2-channel OBRIRs) using the `multiconvolve~` object in the HISSTools Impulse Response Toolbox [13] for Max, which is an implementation of the variable-size partitioned convolution scheme proposed by Gardner [14]. This convolution algorithm offers a good compromise between low latency, IR length and computational cost. Because convolutions are CPU-intense operations, the `multiconvolve~` objects were embedded into `poly~` objects, which distribute the computational load evenly among different CPU cores. The motivation for having all convolutions at different orientations running in parallel is that the user can move the head rather quickly and the system must already be convolving the sound for the next orientation.

The output of Max, which corresponds to the reflections from the VAE at the orientation of the user

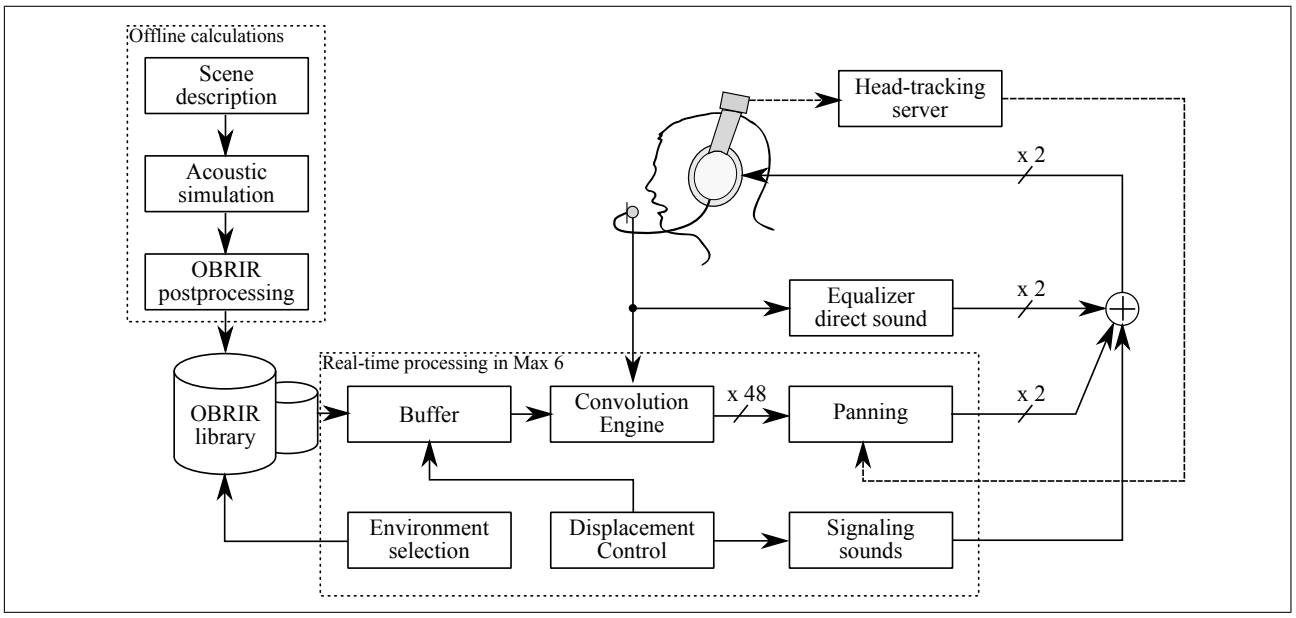

Figure 1. Block diagram of the system for interactive auralization of self-generated oral sounds

in the horizontal plane (yaw), is obtained by a linear panning of the output from the two two-channel convolutions with the OBRIRs which are closest to the current yaw of the user. The output gain of all other convolutions is set to zero, but they remain running continuously so as to be available when the user rotates the head. The orientation of the user is received via Open Sound Control protocol from the head-tracking server, i.e. the computer to which the head tracker is connected via USB.

Finally, the interpolated output, containing the reflections of the VAE, the direct sound compensation from the equalizer, and signaling sounds, which give feedback to the user about their actions, are combined and sent to the user over headphones.

2.2. OBRIR determination

OBRIRs were obtained by acoustic simulation with CATT Acoustic v9.0c. A scene is defined with a geometrical model of the room and the obstacles in it, the absorption and scattering properties of the materials, the placement and radiation characteristics of the source and the placement and orientation of the receiver, together with a Head-Related Transfer Function (HRTF) set that defines the filtering of head and ears to sounds arriving at different angles around the listener. We chose the HRTF set measured by the Institut für Technische Akustik, RWTH Aachen from their artificial head ITA Kunstkopf, which is readily available in CATT v9. The receiver was placed at a height of 1.6 m above the floor and was pointing towards the source, which was located 10 cm away at the same height and pointing away from the receiver. For one receiver location, 24 orientations are calculated, rotating the source around the receiver by 15° and changing the orientation of the receiver accordingly.

It should be noted that the HRTF used was measured at a distance of 2 m. As such, acoustic reflections from surfaces closer than this distance may not be correctly rendered binaurally due to differences in near-field and far-field HRTFs [15].

In CATT Acoustic, the computation parameters were adjusted to automatically choose a suitable number of rays and length of the impulse response, and the second built-in calculation algorithm was chosen. Using MATLAB, the OBRIRs were upsampled to 96 kHz and trimmed to remove the direct sound and the initial delay of the response, in order to account for the latency introduced by the real-time processing of the system. As an alternative to the use of simulated OBRIRs, acoustic measurements of OBRIRs, following the method of Cabrera et al. [7], can be used in the present system.

2.3. Headphone Insertion Loss compensation

A headphone system offers a shorter latency than a loudspeaker system (e.g. [12]), due to the longer propagation time in the latter; nevertheless, headphones alter the airborne direct sound propagation between the mouth and the ears. Open headphones Sennheiser HD570 are chosen to minimize the occlusion effect [16], but nevertheless cause a noticeable insertion loss (IL) at high frequencies and a modification of the resonances in the ear cavity at medium frequencies.

In order to measure and compensate for this IL, a custom dummy head, with a loudspeaker at its mouth and microphones at its ears was used. The SPL at the unobstructed ears, while the loudspeaker at the

mouth was playing a stationary signal (filtered pink noise), was taken as a reference. After placing the headphones on the dummy head, and using only the equalizer and mixer in Fig. 1, the 1/3rd octave band SPLs at the ears were monitored. The gain controls of each of the bands of the EQ were adjusted until the monitored SPLs were closest to the reference SPLs.

A hardware equalizer was preferred to a software-based system because, according to Pörschmann [8], a delay longer than 0.67 ms in the restoration of the direct sound produces noticeable artifacts.

2.4. Latency

The actual latency of the system was determined by placing the auralization system on the head of a user, wearing also miniature microphones at the entrance of the ear canals, and recording the sound produced by a user's oral click when a Dirac delta (without delay) is loaded on the convolvers. The latency is the delay between the direct sound and the processed sound delivered via headphones. According to the measurement at the left ear, shown in Fig. 2, the latency is 3.5 ms.

Such latency allows the simulation of reflections from objects that are more than 70 cm away from the user, considering that the sound propagates at a speed of 340 m/s.

2.5. Calibration

The aim of the calibration is to match the level of the reflections, relative to the direct sound, reproduced by the system to those occurring in reality.

Unlike Yadav [11], who matched the measured OBRIRs in the actual environment to the OBRIRs measured with the auralization system in use, our OBRIRs are obtained with a simulation method that does not correctly represent the airborne direct sound propagation from the mouth to the ears.

For this reason, we chose to define a scene which we could both measure and simulate. In this scene, the source/receiver was placed at 1.5 m in front of a large and totally reflecting wall, pointing towards it. For the measurements in the real scene, a dummy head was placed with the center of the ears 1.5 m above a reflecting plane in a semi-anechoic room, as shown in Fig. 3. The resulting OBRIR (for the left ear) is shown in Fig. 4(a).

For the acoustic simulation, a large rectangular room was modeled, in which only one wall (of dimensions 6 m x 6 m) was totally reflecting while all the other ones were totally absorbing. The receiver was placed at 1.5 m in front of the reflecting wall, at its center, and the source was placed 10 cm in front of the receiver, pointing towards the wall. The OBRIR was obtained and the first 5 ms were set to zero to remove the direct sound from the simulation. This response was upsampled to 96 kHz and used as the kernel for the convolution in Max.

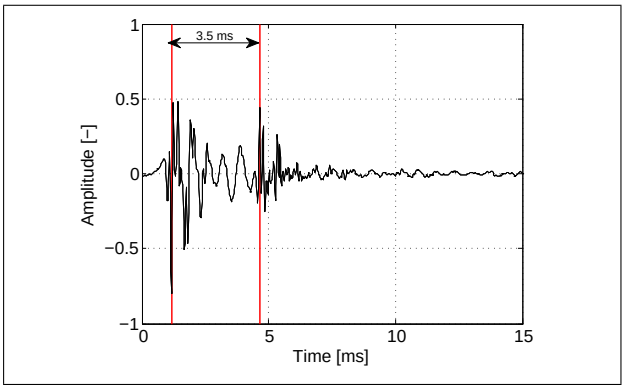

Figure 2. Oral click recorded at the left ear when the convolver is running a Dirac delta.

Figure 3. Dummy head with the center of the ears 1.5 m above a reflecting plane in a semi-anechoic chamber

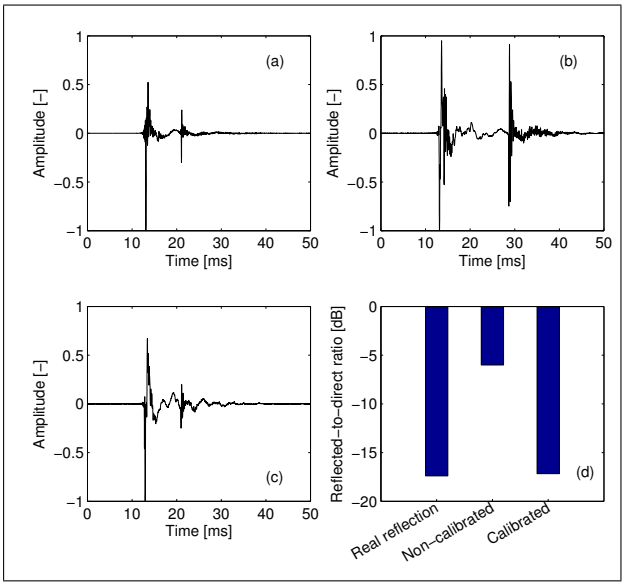

Figure 4. Measured OBRIRs (a) in front of a reflecting plane, (b) with a simulated plane, non-calibrated. (c) with a simulated plane, calibrated. (d) Reflected-to-direct sound energy ratios.

The headphones, microphone, and head tracker were placed on top of the dummy head, and the OBRIR was measured while the auralization system was running (see Fig. 4(b)). The difference in the arrival time of the first reflection relative to the direct sound (see Figs. 4(a) and (b)) is due to observed zero pre-padding in CATT Acoustic. This zero pre-padding is trimmed in order to correctly align the reflected response to the direct sound path via the EQ processor response. In addition, a gain was applied to the simulated OBRIR determined from the variation in reflected-to-direct sound level difference in both measurements (Fig. 4(d)). After applying the gain and trimming the simulated OBRIR used in the convolution, we measured the OBRIR again. The result is shown in Fig. 4(c).

2.6. Virtual environment navigation

Basic environment navigation can be achieved by switching the sets of pre-calculated OBRIRs in real-time.

For enabling this feature, the 24 OBRIR calculations (corresponding to all angles) are repeated for different positions of source-receiver, in a rectangular grid with a spacing of 0.75 m between points, which corresponds to the length of an average footstep.

With the direction cursors on a Wii remote controller, the user can switch positions within the VAE. By choosing a position, Max loads the OBRIRs for the 24 head orientations (at 15° intervals) at that position into a buffer and transfers them to the convolution engine. Additional sounds, like footsteps or verbal information, are played back in order to give feedback to the user.

2.7. Bottlenecks

The current implementation of our system is limited in different aspects that can be improved in future versions.

We chose an angular resolution of 15°, resulting in 24x2-channel simultaneous convolutions that take up most of the CPU processing power. With a finer resolution, our CPU did not have enough processing power to deliver the output of all convolutions in real-time and produced audible clicks. For this reason, displacements are done between discrete positions – and not in a continuous way, as this would require further concurrent convolutions and interpolation.

In dynamic scenarios, when the user rotates the head, the OBRIR that should be applied contains a combination of the receiver orientation in the current orientation of the user and a time-varying source position/orientation corresponding to earlier orientations of the user. This would require further computations in CATT Acoustic for all possible source / receiver orientation combinations. We have made the approximation that the user hears what has been emitted from his current orientation.

The HRTFs used in our system were non-individualized. This can degrade externalization and sound localization performance; nevertheless, the use of head tracking partly accounts for it. Since HRTFs were not individualized, and the used headphones had a fairly flat frequency response, no headphone equalization was performed.

Moreover, the correction of the direct sound has been done in magnitude only, disregarding phase corrections, and for the dummy head. Therefore, some users might perceive noticeable coloration at some frequencies.

3. Concluding remarks

In order to further study and understand human echolocation skills and training, a headphone-based system for interactive auralization of self-generated oral sounds, picked up with a microphone, was implemented. This system delivers the acoustic reflections of these sounds in a virtual environment, by convolution with Oral-Binaural Room Impulse Responses generated via acoustic simulation. The latency of the system is 3.5 ms, allowing the simulation of boundaries as close to the user as 70 cm. By means of head-tracking, the virtual environment remains stable with respect to head rotations of the user, and delivers back the reflections of sounds that would originate from his current orientation.

Initial experiences with the system by blind expert echolocators show that, while the system reacts smoothly and has a pleasant sound, it does not reproduce accurately the sensations of sound reflections occurring at close distances (closer than 3 m), but it provides helpful cues for the discrimination of distant walls or corners (beyond 3 m). The increased difficulty in detecting near reflections asks for further investigations in terms of calibration accuracy, near-field versus far-field HRTFs, individualization of HRTFs, or artifacts in the OBRIRs.

Acknowledgement

This project has been funded by Research Foundation – Flanders (FWO). Portions of the study were carried out during a visiting research stay by the first author at the LIMSI-CNRS.

References

[1] World Health Organization, "WHO | 10 facts about blindness and visual impairment," 2012.

[2] N. Fletcher, "Animal Bioacoustics," in *Springer Handbook of Acoustics* (T. D. Rossing, ed.), ch. 19, pp. 785–804, New York, NY: Springer, 2007.

[3] L. Thaler, "Echolocation may have real-life advantages for blind people: an analysis of survey data.," *Frontiers in physiology*, vol. 4, p. 98, Jan. 2013.

[4] D. Kish and H. Bleier, "Echolocation : What it is, and how it can be taught and learned," tech. rep., Riverside, CA, 2000.

[5] A. J. Kolarik, S. Cirstea, S. Pardhan, and B. C. J. Moore, "A summary of research investigating echolocation abilities of blind and sighted humans.," *Hearing research*, vol. 310C, pp. 60–68, Apr. 2014.

[6] T. Papadopoulos, D. Edwards, D. Rowan, and R. Allen, "Identification of auditory cues utilized in human echolocation: Objective measurement results," *Biomedical Signal Processing And Control*, vol. 6, pp. 280–290, July 2011.

[7] D. Cabrera, H. Sato, W. Martens, and D. Lee, "Binaural measurement and simulation of the room acoustical response from a person's mouth to their ears," *Acoustics Australia*, vol. 37, no. 3, pp. 98–103, 2009.

[8] C. Pörschmann, "One's own voice in auditory virtual environments," *Acta Acustica united with Acustica*, vol. 87, pp. 378–388, 2001.

[9] F. Wefers and M. Vorlander, "Interactive acoustic feedback into virtual acoustic scenes," in *Proceedings of EAA EUROREGIO 2010*, (Ljubljana), 2010.

[10] L. Picinali, A. Afonso, M. Denis, and B. F. Katz, "Exploration of architectural spaces by the blind using virtual auditory reality for the construction of spatial knowledge," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 393–407, 2014.

[11] M. Yadav, D. Cabrera, and W. Martens, "A system for simulating room acoustical environments for one's own voice," *Applied Acoustics*, vol. 73, pp. 409–414, Apr. 2012.

[12] D. Pelegrín-García and J. Brunskog, "Speakers' comfort and voice level variation in classrooms: Laboratory research," *The Journal of the Acoustical Society of America*, vol. 132, pp. 249–260, July 2012.

[13] A. Harker and P. A. Tremblay, "The HISSTools Impulse Response Toolbox: Convolution for the Masses," in *ICMC 2012: Non-cochlear Sound*, pp. 148–155, The International Computer Music Association, 2012.

[14] W. G. Gardner, "Efficient Convolution without Input/Output Delay," *Journal of the Audio Engineering Society*, vol. 43, pp. 127–136, Nov. 1995.

[15] T. Lentz, *Binaural technology for virtual reality*. Berlin, Germany: Logos Verlag, 2007.

[16] M. A. Fagelson, "The Occlusion Effect and Ear Canal Sound Pressure Level," *American Journal of Audiology*, vol. 7, pp. 50–54, Oct. 1998.